Rater Methodology for Stroboscopy: A Systematic Review

*,†Heather Shaw Bonilha, *Kendrea L. Focht, and *,†Bonnie Martin-Harris, *†*Charleston, South Carolina*

Summary: Objectives. Laryngeal endoscopy with stroboscopy (LES) remains the clinical gold standard for assessing vocal fold function. LES is used to evaluate the efficacy of voice treatments in research studies and clinical practice. LES as a voice treatment outcome tool is only as good as the clinician interpreting the recordings. Research using LES as a treatment outcome measure should be evaluated based on rater methodology and reliability. The purpose of this literature review was to evaluate the rater-related methodology from studies that use stroboscopic findings as voice treatment outcome measures.

Study Design. Systematic literature review.

Methods. Computerized journal databases were searched for relevant articles using terms: stroboscopy and treatment. Eligible articles were categorized and evaluated for the use of rater-related methodology, reporting of number of raters, types of raters, blinding, and rater reliability.

Results. Of the 738 articles reviewed, 80 articles met inclusion criteria. More than one-third of the studies included in the review did not report the number of raters who participated in the study. Eleven studies reported results of rater reliability analysis with only two studies reporting good inter- and intrarater reliability.

Conclusion. The comparability and use of results from treatment studies that use LES are limited by a lack of rigor in rater methodology and variable, mostly poor, inter- and intrarater reliability. To improve our ability to evaluate and use the findings from voice treatment studies that use LES features as outcome measures, greater consistency of reporting rater methodology characteristics across studies and improved rater reliability is needed.

Key Words: Voice–Stroboscopy–Reliability–Rater.

INTRODUCTION

It has been estimated that at any given time between 6.6% and 7.5% of the people in the United States have a voice disorder.[1] The cost of voice disorders is estimated to be between 0.7 and 4.9 billion dollars.[2] There are several types of treatments for voice disorders; broad categories of treatments are pharmaceutical, surgical, and behavioral. Treatment outcome measures are used to evaluate the efficacy of these treatments in research studies and clinical practice. There are several different types of outcome measures available to evaluate the effectiveness of a treatment for voice disorders including: patient report, perceptual assessment, acoustic analysis, aerodynamic measures, and laryngeal imaging.[3,4] This article will focus on methodology related to using laryngeal imaging, specifically stroboscopy, as a voice treatment outcome measure.

Three studies epitomize the importance of laryngeal endoscopy with stroboscopy (LES). Sataloff et al[5] studied the clinical value of LES, beyond that provided by clinical and laryngeal mirror examination, and found that LES added diagnostically relevant information in 47% of the cases and that clinically significant findings were detected only through LES in 32.4% of patient cases. These results led the authors to conclude that "stroboscopy is invaluable in daily practice and essential for valid, reliable diagnosis of voice disorders." Similarly, Remacle[6] found that in 732 patients LES findings were considered useful in 92% of cases. Behrman[7] found that 94% of speech-language pathologists who treated patients with voice disorders considered LES important for defining overall therapy goals. In addition, they found that 89% considered LES informative for outcomes assessment, and 81% considered LES important for educating patients about voice production. Although there are some limitations to LES, such as its reliance on pitch tracking and temporal resolution, it remains invaluable for assessing vocal fold structure and function.

LES as a voice treatment outcome tool is only as good as the clinician interpreting the LES recordings. The interpretation and value of stroboscopic findings are directly linked to the training and skills of the operator. Thus, studies using LES as outcome measures rely heavily on their raters. There are a number of rater-related characteristics that should be considered when using LES as an outcome measure including: number of raters, profession of raters (Otorhinolaryngologists [ENT]/ Speech-Language Pathologists [SLP]), blinding, training, use of randomization, interrater reliability, and intrarater reliability. Given the value of LES and the importance of reporting consistent voice treatment outcome measures, we undertook a literature review on these topics. The purpose of this literature review was to evaluate the rater-related methodology from studies that use stroboscopic findings as voice treatment outcome measures.

METHOD

Search strategy

This review was conducted using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement standards.[8] A search strategy was developed and implemented in three computerized journal databases (PubMed, Ovid, and

Accepted for publication June 26, 2014.

Portions of this study were presented at the 43rd Symposium of The Voice Foundation: Care of the Professional Voice. Philadelphia, Pennsylvania, May 2014.

From the *Department of Health Science and Research, Medical University of South Carolina, Charleston, South Carolina; and the †Department of Otolaryngology—Head and Neck Surgery, Medical University of South Carolina, Charleston, South Carolina.

Address correspondence and reprint requests to Heather Shaw Bonilha, Department of Health Science and Research, Medical University of South Carolina, 77 President St, Charleston, SC 29425. E-mail: bonilhah@musc.edu

Journal of Voice, Vol. 29, No. 1, pp. 101-108

0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.06.014

Cochrane) to identify all English language studies where LES was used as an outcome measure for the treatment of a voice disorder. The following search terms were used: "laryngostroboscopy," "stroboscopy," "strobovideolaryngoscopy," "strobolaryngoscopy," "videostroboscopy," and "videolaryngostroboscopy." Each of the search terms was combined with "treatment." Specifically, the PubMed search was: laryngostroboscopy, stroboscopy, strobovideolaryngoscopy, strobolaryngoscopy, videostroboscopy or videolaryngostroboscopy and treatment. All studies published from database inception (PubMed and Ovid electronic 1946, Cochrane 1993) to our last search on November 21, 2013 were reviewed for eligibility. Unpublished reports were not considered for this review. Authors were not contacted.

Inclusion criteria

One reviewer (either K.L.F. or H.S.B.) assessed each study based on the following "inclusion" criteria in the following order: English language; original article; human study; perceptual judgments of stroboscopic findings reported both pretreatment and posttreatment; and aggregate data reported for five or more participants. Duplicate results were deleted. Then, a second reviewer (either K.L.F. or H.S.B.) assessed each study identified by the first reviewer for inclusion criteria.

Assessment of evidence

Eligible articles included in this review were categorized and evaluated for the use of rater-related methodology, reporting the number of raters, types of raters, blinding, and rater reliability by H.S.B. and K.L.F.

Data synthesis

The analysis was descriptive in nature because the heterogeneity of rater methodologies and data used to report rater reliability precluded a robust statistical analysis (ie, meta-analysis).

RESULTS

Assessment of evidence

Of the 738 articles reviewed, 80 articles met inclusion criteria (see Figure 1). Eligible studies are summarized in Table 1.

Number and types of raters

More than one-third of the studies included in the review (30/80, 38%) did not report the number of raters who participated in the study. Of the 50 studies that did report the number of raters, nine (18%) used one rater, 20 (40%) used two raters, 16 (32%) used three raters, four (8%) used four raters, and one (2%) used six raters. Forty-six of the 80 (58%) articles specified the profession of the raters. The raters were ENTs (otolaryngologists, phoniatricians, and laryngologists) in 32 articles, both ENTs and SLPs in 12 articles, and SLPs only in two articles.

Rater blinding and recording randomization

Twenty-five of the 80 studies (31%) reported using raters blinded to treatment status. Sixteen out of the 80 studies (20%) reported randomizing the LES recordings for rating.

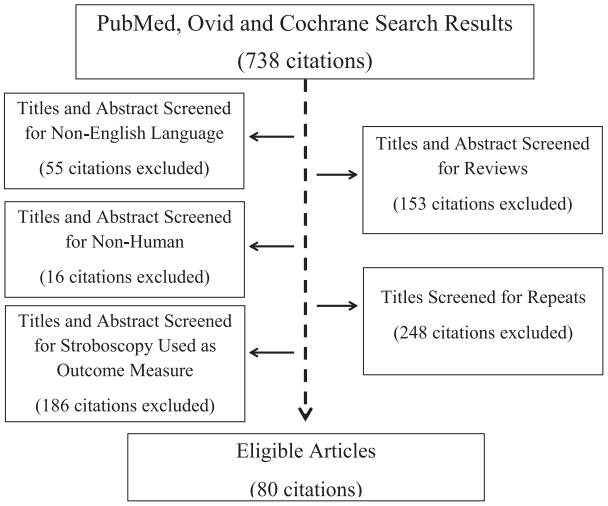

FIGURE 1. Systematic review flow diagram.

Rater reliability reporting

Eleven studies (14%) reported results of rater reliability analysis. Six articles reported results for interrater and intrarater reliability, four reported only intrarater reliability and one reported only interrater reliability (Table 2). Fifteen of the 80 articles reported the use of consensus rating. One article, Galletti,[9] reported highly concordant results between raters but did not report the methods used to test rater reliability or results.

Intrarater reliability results. In the 10 studies that reported intrarater reliability, five studies reported correlation coefficients, two reported exact percent agreement, two reported κ values, and one study reported correlation and regression slope values. Correlation coefficients ranged between 0.17 and 0.93. Percent agreement values ranged from 0 to 100, and κ values ranged from 0.098 to 1.0.

Five of the 10 studies (50%) reported good intrarater reliability (per study-specific criteria in Table 2). Lam's[10] study included data from 82 patients who underwent LES examinations at four time points: one before, two during, and 1 after a treatment/placebo period. Lam reported the highest intrarater reliability at above 0.90, although, this was based solely on one examiner rerating eight recordings. Furthermore, the features rated by Lam were from the Reflux Finding Score,[11] which are all anatomical and stationary. Karpenko[12] reported high intrarater reliability (89%) for ratings of supraglottic activity, mucosal wave, and glottal competency using a 5-point scale. However, only four LES recordings were rated to obtain intrarater reliability data in Karpenko study. Because good intrarater reliability was found for four consecutive LES recordings, only one rater scored the remaining six recordings. Wang[13] reported good intrarater reliability (0.80) from 10% of their recordings rerated simultaneously by two ENTs using a consensus method. This was the only study that used a consensus scoring approach and reported reliability results. Beaver[14] reported correlation coefficients indicating a good level of agreement, although their data were based on rating elements such as, edema or erythema using a 4-point scale rather